CODIB-D-39
1 July 1959

UNITED STATES INTELLIGENCE BOARD

COMMITTEE ON DOCUMENTATION

MEMORANDUM FOR: Committee on Documentation

SUBJECT: A Proposed Notation Scheme for the Intelligence Subject Code

25X1

The attached study by [redacted] is circulated for your information. It was prepared under a contract between [redacted] and the Intelligence Laboratory, Directorate of Intelligence and Electronic Warfare, RADC, and is now before CODIB's Working Group on the Revision of the ISC for discussion and possible adoption. Comments and suggestions on this proposal should be submitted to your representatives on the Working Group prior to its next meeting on 9 July 1959.

25X1

25X1

[redacted]
Secretary

25X1

CONFIDENTIAL

A PROPOSED NOTATION SCHEME FOR THE INTELLIGENCE SUBJECT CODE

INTRODUCTION

As an extension of work that it is doing for the Rome Air Development Center and other organizations in the Intelligence Community in the design and improvement of information storage and retrieval systems, the firm of [redacted] was requested by the Committee on Documentation of the United States Intelligence Board to develop a notation system for the newly-revised Intelligence Subject Code. The brief report that follows is an account of the thinking that led to the notation that was produced. 25X1

Stated very generally, there were three basic requirements to be met in the design of a Code notation. The first was broad assessibility to the various storage and retrieval systems in present and potential use within the Intelligence Community. The second was expansibility to provide for future changes in the subject interests of the several agencies that might use the Code. The third requirement was for facility of use both in coding and in searching.

Basic to the considerations that led to the notation that was developed was the fact that it had been generally agreed that the Intelligence Subject Code was, in the future, to be used for the subject coding of information contained in intelligence documents, and not as a basis for the shelving of books in intelligence libraries. Thus, the notation is not designed from the viewpoint of shelf or file arrangements, but, rather, as a subject basis for non-conventional storage and retrieval systems utilizing optical, electromechanical, and other searching devices.

THE CHARACTERS IN THE NOTATION

Because at least two members of the Intelligence Community planned to use it in a storage and retrieval system utilizing Minicards, it was decided that the notation should have six characters. Minicards require a six character code, which can be numerical, alpha-numeric, or alphabetical. (Actually, the reason that the Minicard system was designed for a six character code is that the present Intelligence Subject Code has a six character numerical notation. And so it would perhaps be just as correct to say that the reason the revised Code is to have a six character notation is that the present one does.)

Alpha-Numeric vs. Numeric Notation

Early in their deliberations, the designers of the revised Code gave serious consideration to the use of an alpha-numeric notation. The rationale behind this was two-fold. First, since there are 26 letters as opposed to only 10 Arabic numerals, it becomes possible to produce many more discrete combinations with letters than with numbers. This makes it easier to accommodate the various classes and levels of classes within the Code, and to

provide for expansion. Second, it was felt that there is a smaller likelihood of errors in the transcription of an alpha-numeric code than there is in the transcription of a purely alphabetic or purely numeric code.

The problem that led to the need for the very large number of discrete combinations that an alpha-numeric notation could provide arose from a general desire for a hierarchical notation, with its distinct mnemonic advantages for both the coder and the searcher. The main drawback of hierarchical notations is that they are very wasteful of notation characters. An alpha-numeric code, with its great combination potential, provided ample insurance against ever running out of discrete notations, even with a hierarchical code.

Unfortunately, another overriding consideration made the use of an alpha-numeric notation impractical. This was concerned with the imminent use of the Intelligence Subject Code in storage and retrieval systems utilizing binary notations to represent decimal numbers and letters of the alphabet. Where binary numbers are used to represent characters in a subject code, there is a significant element of economy in using a numerical rather than an alphabetic or alpha-numeric notation. This arises from the fact that it takes far fewer binary digits or "bits" to represent a decimal number than it does to represent a letter of the alphabet. Since binary reading mechanisms are capable of scanning only a finite number of "bits" in a given interval, the smaller the number of "bits" per code character the faster the scanning or searching rate. Thus, from the viewpoint of search speed, a numerical notation is more economic than an alphabetic or alpha-numeric notation, although the latter do permit a greater number of discrete combinations and are therefore more amenable to hierarchical arrangement. As it finally resolved itself, the essential purpose of the study upon which the present report is based was to develop a notation system combining the economies of numerical characters with the mnemonic advantages of a hierarchic arrangement.

THE STRUCTURE OF THE NOTATION

There are between 5000 and 6000 subject classes in the revised Intelligence Subject Code. If used consecutively, without hierarchy, a six-digit notation beginning with six zero's and ending with six nine's (or vice versa) could easily accommodate all of the existing subject classes, and leave space for practically indefinite expansion. However, when hierarchy is injected into the system, the picture changes drastically.

Inapplicability of Pure Hierarchy

Beginning with the chapter numbers and working downward, there are a total of six hierarchic levels in the revised Code. If one were to assign one digit in the six-digit notation to each level, it would seem theoretically possible to accommodate both strict hierarchy and a six-digit numerical

notation. But the use of such an arrangement is complicated by the fact that it permits a maximum of only ten classes on any level. Thus, suppose the first digit in the notation is dedicated to the chapter number, and there are eleven chapters. There would be no room for the eleventh chapter.

Actually, in the case of the revised Code, the pressure did not arise at the chapter level. It first occurred at the level immediately below the chapters, at the "second digit" level. Two of the chapters had eleven sub-parts, meaning that they could not be accommodated by single digits. Then, on the next level (the third digit), there was one case of a class with twelve sub-parts and three cases where there were eleven. At the "fourth digit" level, there were nine cases where classes had more than ten sub-parts; in fact, one had nineteen. At the "fifth digit" level, there were twenty-three classes with over ten sub-categories. One of these classes had thirty-three sub-parts. On the sixth and final level, there were four classes having over ten sub-categories.

From the foregoing analysis, it is clear that pure hierarchy, with each level being given equal value, could not be applied to the present situation. Aside from the most general level, each level in the hierarchy has several classes containing more than ten sub-categories, making any single digit per level arrangement impossible, in view of the limitation of six characters. With strict hierarchy, an eleven-character notation would be required, allowing one digit for the first level and two for each of the remaining five.

Use of a Modified Hierarchy

The problem of overcrowding was finally remedied by not applying notations to the most general levels of the Code, which were used merely to identify chapters and broad subdivisions within chapters and had a very small likelihood of being searched. In practice, it developed that this meant merely skipping the second highest level of generality, since each chapter in the Code was given a different initial digit. Thus, as can be seen from Appendix I, by searching under code "500000," it is possible to obtain all materials in a collection on the productive capacity of a given country. It would also be possible to do generic searches on the second highest level of generality (Agriculture, Forestry, and Fisheries, etc., in Appendix I) by making this level coordinate or equal to the third level for the purposes of the notation. However, this would mean that each time an item on the third level (such as Processed Animal Products in Appendix I) is coded a code entry would also have to be made for the second level (Agriculture, Forestry, Fisheries, etc. in this case).

Aside from the identification of each Code chapter with a different initial digit in the notation, mnemonic assistance to the searcher and coder

is derived from the fact that, beginning with the fourth level of generality (the fourth digit in the notation), there is a strict hierarchical breakdown. As will be noted in Appendix II, the code number "507400" is given to Hides and Skins, "507410" is given to Leather and Leather Products, and "507411" is given to Leather Apparel. This facilitates the rapid identification of the notations applied to a given subject, and it helps to minimize coding and transcription errors.

Returning to Appendix I, it will be noted that several of the major subject classes have two six-digit numbers applied to them. For instance, Organic Chemicals in numbered "516000" and "517000." The reason for this is that this subject is so large that it could not be accommodated within a single major class. However, this necessary compromise does not seriously diminish the mnemonic qualities of the system; it merely means that the coders and searchers have to remember "516" and "517" rather than only "516" in dealing with materials on Organic Chemicals.

CONCLUSIONS

In evaluating and using the proposed notation scheme, it should be borne in mind that its effective life is probably no more than ten years, and very likely considerably less. While a very conscious effort was made to provide for expansibility, an appreciable sacrifice of notation characters was made in order to provide an element of hierarchy within the bounds of the available six digits. However, wherever possible and indicated, unused code numbers were left, to provide for future expansion. Thus, in Appendix I, every fourth code number is left blank in order to permit the inclusion of new subjects in proximity to old subjects most closely related to them. In Appendix II, the category, Dairy Products and Eggs, ends with the code number, "507250," and the next category, Fish Products, begins with the code number, "507300." In addition to providing for future expansion, this interval exploits the mnemonic advantages of having the major classes clearly separated.

One major area of consideration has been omitted from the present discussion. This is concerned with the Subject Modifiers that occur within the various chapters of the Code. The reason that the subject of modifiers has been omitted is that it has been decided that they should be treated in a code schedule quite separate from that presently under discussion. It has also been suggested, with good reason, that every effort be made to ensure that the modifiers do not overlap subject classes that are already in the body of the Subject Code. Another suggestion is that careful consideration be given to the inclusion of highly specific modifiers as subjects in the body of the Code. The modifiers that remain, and that are treated in the separate code schedule, should then be those that have, or could have, truly universal application.

Regarding the form or arrangement of the modifiers, several suggestions have been put forward. One is to simply list them alphabetically, with adequate definitions or "scope notes," and to permit the coders to use them whenever a required qualifying detail about a subject is not present in the body of the Code. This has the merit of flexibility, but it also presents the danger of misuse by the coders. This danger might perhaps be alleviated through the "scope notes" and by means of a carefully-worded set of rules as to when and how modifiers are to be used. If the straight alphabetical arrangement were adopted, a three-digit notation could be used, with very adequate space for the addition of new modifiers in the future.

Another possible treatment of the modifiers is to make a schedule similar to the Intelligence Subject Code itself, with different chapters dedicated to modifiers dealing with different subjects. With this arrangement, modifiers of a truly universal nature, which could be applied to any and all subjects, would be put in a chapter given to general modifiers, and those modifiers that are applicable only to one or several specific subjects would appear in chapters of the "modifier schedule" dealing with the subject or subjects involved. This arrangement has the advantage of putting all of the modifiers that may be used for a given subject in one place. This can serve as a mnemonic aid to the coder.

As for the notation for the modifiers under this second arrangement, two possibilities commend themselves. One is to give a modifier the same notation in all chapters of the "modifier schedule" in which it appears. This would constitute a classified version of the straight alphabetical listing of the modifiers, and the same three-digit notation would be used. The second possibility would be to give the modifiers in each chapter of the "modifier schedule" a different schedule. This would mean that a given modifier, applied to several subjects, would have a different notation for each subject to which it is applied. This approach has a possible mnemonic advantage for the coder.

Another phase of the overall coding problem that was omitted from the body of this report has to do with the Area Classification Code. Again, this omission was intentional, since the Area Classification Code is quite separate from the Intelligence Subject Code. However, geography is basic to any storage and retrieval system involving intelligence materials, and due attention has been given to this phase of the problem.

In general, the Area Classification Code, with its six letter notation which encompasses hierarchy, appears completely satisfactory from both the coding and searching viewpoint (aside from the previously mentioned fact that letters of the alphabet are not as economic of "bits" as decimal numbers). The only two things that appear to be missing from the Area Classification Code are a clear-cut description of how it should be used, and a breakdown which is finer than regions within a country.

The first omission speaks for itself. In regard to the second, we would suggest that a separate code schedule be made for cities. Whether this is done on a classified basis by country or on a straight non-hierarchical basis, such a schedule of cities could easily be encompassed in a six-digit numerical notation. The use of such a separate schedule would be very useful in retrieving information about specific cities. At present, this is done indirectly, and is rather difficult.

Next 1 Page(s) In Document Denied